Detecting CpG islands using Hidden Markov Model

Jihoon Kim
(2002-23331)

SNUBI:SNUBiomedical Informatics,
Seoul National University College of Medicine

Introduction

- A Basic HMM will be designed to find the CpG islands. Prototype is to be implemented using Matlab followed by the real simulation of C-program. Evaluation will be held by counting f True and False values applied to the EMBL CpG database.

What is a CpG island?

- General definition:
  CpG island is a short and dispersed region of unmethylated DNA with a high frequency of CpG dinucleotides relative to the bulk genome
- Gardiner-Gardin's definition: *Gardiner-Garder(1987)*
  regions of DNA satisfying the followings.
  - Length : at least 200 bp
  - G+C content: above 50%
  - Ratio of observed vs. expected: above 0.6

Why CpG island?

- CpG islands are useful markers for genes in organisms containing 5-methylcytosine in their genomes.
- CpG islands may be sites of interaction between transcription factors and promoters
- Methylation of promoter CpG islands plays an important role in *Takai(2001)*
  - gene silencing
  - genomic imprinting
  - X-chromosome inactivation
  - the silencing of intragenomic parasite
  - and carcinogenesis

Dataset

- Human CpG-island database 4.0
  - EMBL(European Molecular Biology Lab.) nucleotide sequence database
  - Among 1711 entries,
    1211 with CpG islands
    499 without CpG island
  - Larsen et al.(1992)
  - ftp://ftp.no.embnet.org/cpgisle
- GeneBank : for sequence
- Randomly select from the cpgisle database
  - Training data
  - Test data

CpGisle - excerpt

```
ID   GAPDHG
AC   J04038;
LE   5378
DE   Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene, complete cds.
DE   7/95
EX   Gene expression widespread
FT   CpG island      871..1673
FT                   /size=803
FT                   /%(C+G)=69.12
FT                   /Obs/Exp CpG=0.82
FT   CpG island      1683..2063
FT                   /size=381
FT                   /%(C+G)=67.19
FT                   /Obs/Exp CpG=0.77
XX
FT                   /CAAT-box.1="884"
FT                   /CAAT-box.2_complement="2156"
FT                   /GC-box="1064"
FT                   /E2F_CS.1="1785"
FT                   /SpI="158,1198,1244,1290,1310,1314"
FT                   /SpI_complement="174,584,1519,1668,1736,2271"
FT                   /SpI_complement="2625"
FT                   /AccII="717,727,1093,1268,1334,1423"
FT                   /AccII="1489,1531,1788,2006,3650,4278"
//
```

Three fundamental problems of HMM

- Evaluate the probability(or likelihood) of a sequence of observation, given a specific HMM
- Determine a best sequence of model states
- Adjust the model parameters so as to account for a set of observed signals as good as possible

Assumption

- At any point of the input sequence, observing a nucleotide is a probabilistic function of two corresponding stochastic processes of a HMM
- Stochastic process I:
  - visible
  - treat DNA sequence as a time ordered sequence, and observing a nucleotide at a position depends probabilistically on the previous nucleotide
- Stochastic process II:
  - invisible
  - any portion of the observed sequence is either in a CpG-island or not
- Hence, there are 8 states in all:
  A^+, B^+, C^+, D^+, A^-, B^-, C^-, D^-

Method

- Find for each positions of the sequence, the longest CpG-island with a probability greater than a certain threshold
- Superpose the results to get a global picture of the sequence
- See if these are consistent with those obtained from published works

Algorithm

- Initialize the probability of having a length(=window size) one CpG island
- Then, seek for CpG islands having length more than one
- At each iteration increase the window size by one and calculate the probability of being or not being a CpG island for each nucleotide position

Implementation

- Viterbi algorithm:
  Implement CpG island HMM to predict the most probable path for the test sequence
- Forward & backward algorithm:
  Implement posterior decoding to make predictions about the locations of CpG islands
- Design a more complex HMM

Program

- Input:
  - Length
  - G+C content
  - Observed CpG / Expected CpG
- Output:
  - Start & end positions
  - Length
  - G+C content
  - Observed CpG / Expected CpG

Tool

- Matlab(for prototype)
  - Simple
  - With many predefined functions for matrix
  - Easy to track values of the variables
- C language(for simulation)

Evaluation

- Calculate accuracy using the numbers of:
  - TP: True Positives
  - TN: True Negatives
  - FP: False Positives
  - FN: False Negatives

Reference

- Bird et al.
  "CpG island as Gene Markers in the Human Genome",
  Trends. Genet.(1987)
- Davuluri et al.,
  "Computational identification of promoters and first exons in the human genome", *Nature Genetics(2001)*
- Durbin et al.,
  "Biological sequence analysis",
  Cambridge University Press (1998)
- Gardiner-Garden et al.
  "CpG islands in vertebrate genomes", *JMB(1987)*
- Takai et al.
  "Comprehensive analysis of CpG islands in human chromosomes 21 and 22", *PNAS(2001)*